GenePattern

PreprocessDataset Documentation

Module name:	PreprocessDataset
Description:	Perform several prerocessing options on a res, gct, or Dataset input file
Author:	Joshua Gould, Pablo Tamayo (Broad Institute), gp-help@broad.mit.edu
Date:	04/04/04

Summary: The PreprocessDataset module performs a variety of pre-processing operations including thresholding/ceiling, variation filter, discretization and normalization:

Thresholding:
Value = threshold if Value < threshold
Value = ceiling if Value > ceiling

Variation filter (exclude rows for which):
max / min < minchange
max – min < mindelta
here the max and min are computed over a row excluding the top (and bottom) “num.excl” experiments. This is to prevent one or more “spikes” to make the gene pass the filter.

The filter flag controls the application of both thresholding and the variation filter.

Independently of the application of thresholding and the variation filter the module also has a preprocessing flag to turn on the discretization or normalization of the dataset (after thresholding and filtering).

Probability threshold allows sampling of the rows without replacement to obtain that fraction of the total number of rows. The “max sigma binning” parameter controls how many bins are used when discretizing. The default “value of 1” produces binary discretization (above and below the mean).

The module also includes the option to take the log base 2 of all values in the input dataset. Lastly, the module can remove rows in which the given number of columns does not contain a value greater or equal to a user defined threshold.

The order of the steps in the module is as follows:
1. Thresholding
2. Log Base 2
3. Remove row if n columns not >= than given threshold
4. Variation filter

Parameters:

Name	**Description**
input.filename:	input filename - .res, .gct, Dataset
output.file:	Output file with preprocessed dataset
output.file.format:	output file format

filter.flag:	Variation filter and thresholding flag
preprocessing.flag	Discretization and normalization flag
minchange:	Minimum fold change for filter
mindelta:	Minimum delta for filter
threshold:	Value for threshold
ceiling:	Value for ceiling
max.sigma.binning:	Maximum sigma for binning
prob.thres:	Value for uniform probability threshold filter
num.excl:	Number of experiments to exclude (max & min) before applying variation filter
log.base.two	Whether to take the log base two after thresholding
number.of.columns above.threshold	Remove row if n columns not >= than given threshold
column.threshold	Threshold for removing rows

Return Value:

the filtered, preprocessed output file

Platform dependencies:

Task type:	Preprocess&Utility
CPU type:	any
OS:	any
Java JVM level:	1.4
Language:	Java